

**SPEECH RECOGNITION APPARATUS AND COMPUTER SYSTEM THEREFOR,
SPEECH RECOGNITION METHOD AND PROGRAM
AND RECORDING MEDIUM THEREFOR**

Inventor(s):

Nobuyasu Itoh

Masafumi Nishimura

International Business Machines Corporation

IBM Docket No. JP9-2001-0337 US1

IBM Disclosure No. JP8-2000-1003

Express Mailing Label No. EL 649719607 US

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of Japanese Application No. 2001-044186, filed February 20, 2001.

BACKGROUND OF THE INVENTION

Technical Field

The present invention relates to recognizing the natural speech of a person and converting the speech into text, and more particularly, to automatically removing meaningless words called disfluencies from the derived text.

Description of the Related Art

A statistic method for performing speech recognition using an acoustic model and a language model is disclosed in, for example, "A Maximum Likelihood Approach to Continuous Speech Recognition", L. R. Bahl et al. IEEE Trans. Vol. PAMI-5, No. 2, March 1983; or in "Word-based Approach To Large-vocabulary Continuous Speech Recognition For Japanese", Nishimura, et al., Information Processing Thesis, Vol. 40, No. 4, April, 1999. Further, an N-gram estimate, a common language model technique, is disclosed on page 15 of IBM ViaVoice98 Application Edition (Info-creates Publication Department, issued on September 30, 1998).

Disfluencies, such as "eh", frequently appear during the recognition of natural speech and are important for applications. "Statistical Language Modeling for Speech Disfluencies", by A. Stolcke and E. Shriberg, Proc. of ICASSP96, discloses a method for handling such disfluencies in the N-gram model and automatically removing them from the recognition result. With this method, however, it is difficult to avoid a phenomenon whereby the validity of a word when originally used is not recognized and the word is thereafter determined to be a disfluency and removed. Further, the kind and frequency of disfluencies varies depending on the speaker and the speaking environment (e.g., with or without a draft and in a formal or an informal setting), making it difficult to use an average model for the prediction of disfluencies.

SUMMARY OF THE INVENTION

It is one object of the present invention to provide a function for setting a model that is appropriate for the removal of disfluencies by using information that is designated by a user or that is obtained during the registration of a speaker.

Accordingly, a general-purpose language model as well as a disfluency language model are included as language models to facilitate the automatic removal of disfluencies.

A speech recognition means in wide use today employs a vocabulary/language model (called a topic) that is prepared for a specific field, such as computers, and used in conjunction with a general-purpose language model to improve the recognition accuracy for the pertinent field. A topic having a comparatively small size can be prepared for which setting the ON/OFF state is easy. Thus, according to the present invention, the vocabulary/language model used for removing disfluencies is prepared as a topic.

Specifically, according to the present invention, a dictionary in which disfluencies such as "eh" and "well", as well as normal words, are registered, is prepared along with a Baseform Pool that includes their pronunciations. When disfluencies (<eh> or <well>) do arise, however, to efficaciously and easily distinguish the disfluencies from normal words during a subsequent removal process, a special symbol (e.g., a "not equal" symbol) is used to mark them. Furthermore, a special language model for predicting disfluencies is prepared as a topic. Since the language model for disfluency prediction is used as an element of an N-gram language model including disfluencies, and in the colloquial expressions unique to spontaneous speech, the language model for the disfluency prediction can be much smaller than the general-purpose language model. When this language model is linearly interpolated with the general-purpose language model using the following equation (1), the accuracy for the prediction of a disfluency can be improved.

$$\text{Pr}(w_1|w_2, w_3) = \lambda P_1(w_1|w_2, w_3) + (1-\lambda)P_2(w_1|w_2, w_3) \dots (1)$$

For the example in equation (1), $N=3$ and $0 < \lambda \leq 1$, $P1$ denotes the probability using a general-purpose language model, and $P2$ denotes the probability using the language model (disfluency language model) of a disfluency topic.

The present invention is provided based on the above-described subject matter.

5 A speech recognition apparatus can include a transformation processor for transforming a phoneme sequence or sequences included in speech into a word sequence or sequences, and for providing, for the word sequence or sequences, an appearance probability indicating that the phoneme sequence originally represented the word sequence or sequences; a renewal processor for renewing the appearance probability, provided for the word sequence or sequences by the transformation means, based on a renewed numerical value indicated by language models corresponding to the word sequence or sequences provided by the transformation processor; and a recognition processor for selecting the word sequence, or one of the word sequences for which the renewed appearance probability is the highest, to indicate that the phoneme sequence or sequences originally represented the selected word sequence, and for recognizing the speech. The renewal processor can calculate the renewed numerical value using a first language model, which is especially prepared for expressions unique to spontaneous speech, and a second language model, which differs from the first language model, and employs the renew numerical value to renew
20 the appearance probability.

A disfluency is included in a word set related to an expression unique to spontaneous speech. Therefore, according to the invention, without limiting it to disfluencies, the first language model is included for the expression unique to spontaneous speech. Expressions unique to spontaneous speech are, for example, "I mean" and "you know".
25

In the speech recognition apparatus of the invention, the first language model represents a probability that the word sequence which includes a predetermined word included in an expression unique to spontaneous speech is the word sequence, the phoneme sequence, or sequences originally represented.

According to the speech recognition apparatus of the invention, when the predetermined word is included in the speech recognition results, the transformation processor transforms the phoneme sequence or sequences into a word sequence included in the predetermined word, and the renewing processor renews the appearance probabilities of the word sequence or sequences based on the first language model and the second language model.

According to the invention, the first language model can employ a word set including a disfluency as an element. In the speech recognition apparatus of the invention, the first and the second language models are defined as N-gram models, and the renewal processor can employ, as the renewal numerical value, the weighted average value of the first and the second language models.

In the present invention, the speech recognition apparatus disclosed herein can be implemented in a computer system.

According to the present invention, the following speech recognition method is provided. A speech recognition method according to the present invention can include transforming one or more phoneme sequences included in speech into one or more word sequences, and providing, for the thereby obtained word sequence or sequences, an appearance probability indicating that the phoneme sequence originally represented the one or more word sequences; renewing the appearance probability provided for each of the one or more word sequences when the word sequence or sequences obtained by the transforming step include a word unique to spontaneous speech, referring to a first model, which is especially prepared for an expression unique to spontaneous speech, and a second model, which differs from the first model; and recognizing the speech by selecting the word sequence, or one of the word sequences for which the renewed appearance probability is the highest, to indicate that the phoneme sequence or sequences originally represented the selected word sequence.

In the speech recognition method of the invention, the first language model can be written in correlation with the appearance probability of a word sequence that includes a word unique to spontaneous speech with a combination of N consecutive

words. Additionally, a disfluency is a typical example of the word unique to spontaneous speech.

In the speech recognition method of the invention, the renewing step can include renewing the appearance probability provided for the word sequence or sequences by referring to a third language model, which is especially prepared for specific symbols included in the word sequence or sequences. The specific symbols can include symbols such as a period, a comma, a question mark, and the like. These specific symbols can be automatically inserted.

The speech recognition method of the invention can be established as a program that permits a computer to perform a predetermined process. According to the present invention, a program is provided that permits a computer to acoustically analyze speech data and transform the speech data into a feature vector. The program further can cause the computer to generate acoustic data, for which an appearance probability is provided, for one or more phoneme sequences that may correspond to the feature vector obtained by the acoustic analysis step; transform the phoneme sequence or sequences into one or more word sequences while a disfluency is included as a word choice selection; renew the appearance probability by referring to a disfluency language model that is written by correlating the appearance probability of one or more word sequences in which a disfluency is included with a combination of N consecutive words; and recognizing the speech data using the word sequence, or one of the word sequences, for which the renewed appearance probability is the highest as a speech recognition result.

In the program of the invention, the word transforming step further can add a symbol to a word indicating that the word is a disfluency word choice in order to distinguish the word choice from another word. The symbol is added as a mark that clearly identifies the word as a disfluency, and is also used when a disfluency is to be removed automatically. In the program of the invention, the recognition step can output the word sequence to which the highest appearance probability applies as text data. Before being output, the word to which the symbol has been added is removed so that

the word sequence which has the highest appearance probability can be determined to be text data. This is efficacious for the automatic removal of disfluencies and the display of text data. In the program of the invention, it is practically necessary, at the renewing step, for the appearance probability to be renewed by referring not only to the

5 disfluency language model, but also to a general-purpose language model.

In the program of the invention, at the word transformation step, the phoneme sequence is transformed into a word sequence, while a pause included in the speech data is included as a punctuation choice. At the renewing step, the appearance probability can be renewed by further referring to a punctuation language model that is limited to punctuation insertion. This method is efficacious for the automatical insertion of punctuation.

A speech recognition method of the invention is established also as a storage medium on which a computer-readable program is stored. That is, according to the invention, a storage medium is provided on which a computer-readable program is stored that permits a computer to acoustically analyze speech data and transform the speech data into a feature vector; generate acoustic data for which an appearance probability is provided for a phoneme sequence that may correspond to the feature vector obtained at the acoustic processing step. When a disfluency is to be reflected in a recognition result, the computer-readable storage medium further can cause the

20 computer to transform the phoneme sequence into a word sequence with a disfluency being included as a word choice. When a disfluency is not to be reflected in a recognition result, the phoneme sequence can be transformed into a word sequence without including the disfluency as a word choice.

When a disfluency is to be reflected in a recognition result, the computer-

25 readable storage medium can cause the computer to renew the appearance probability by referring to the general-purpose language model and a disfluency language model, which is written by correlating the appearance probability of the word sequence that includes a disfluency with a combination of N consecutive words. When a disfluency is not to be reflected in a recognition result, the computer-readable storage medium can

cause the computer to renew the appearance probability by referring to the general-purpose language model. The computer-readable storage medium further can cause the computer to recognize the word sequence for which the renewed appearance probability is the highest as a speech recognition result.

5

For the storage medium of the invention, as well as the above described program of the invention, the transforming step can include adding a symbol to a word that is a word choice for a disfluency in order to distinguish between the word and another.

For the storage medium of the invention, the disfluency language model and the general-purpose language model can be N-gram models, and at the renewing step, the appearance probability can be renewed by using the weighted average value of the disfluency language model and the general-purpose language model.

Further, the storage medium of the invention can cause the computer to automatically insert punctuation in a speech recognition result. Accordingly, at the word transformation step the phoneme sequence is transformed into a word sequence, while a pause is included as a punctuation choice in the speech data. At the renewing step, the appearance probability also can be renewed by referring to a punctuation language model that is limited to punctuation insertion.

BRIEF DESCRIPTION OF THE DRAWINGS

There are shown in the drawings, embodiments which are presently preferred, it being understood, however, that the invention is not limited to the precise arrangements and instrumentalities shown.

5 Figure 1 is a diagram showing a configuration of a speech recognition apparatus according to a first embodiment of the invention disclosed herein.

Figure 2 is a diagram showing a configuration of a speech recognition program according to a first embodiment of the invention disclosed herein.

Figure 3A is a diagram showing the operation of the speech recognition program of Figure 2 for speech recognition using only a general-purpose language model.

Figure 3B is a diagram showing the operation of the speech recognition program of Figure 2 for speech recognition using a general-purpose language model and a disfluency language model.

Figure 4 is a flowchart showing the processing performed by the speech recognition program (language decoder) of Figure 2.

Figure 5 is a diagram showing the operation of the speech recognition program (language decoder) of Figure 2 wherein the disfluency language model is not used when performing automatic disfluency deletion (topic OFF).

Figure 6 is a diagram showing the operation of the speech recognition program (language decoder) of Figure 2 wherein the disfluency language model is used when performing automatic disfluency deletion (topic ON).

Figure 7 is a diagram showing a configuration of a speech recognition program according to a second embodiment of the invention disclosed herein.

Figure 8A is a diagram showing the operation of the speech recognition program of Figure 7 for speech recognition using only a general-purpose language model.

Figure 8B is a diagram showing the operation of the speech recognition program of Figure 7 for speech recognition using a general-purpose language model and a disfluency language model.

Figure 8C is a diagram showing the operation of the speech recognition program

of Figure 7 for speech recognition using a general-purpose language model and a punctuation language model.

Figure 8D is a diagram showing the operation of the speech recognition program of Figure 7 for speech recognition using a general-purpose language model, a
5 disfluency language model, and a punctuation language model.

Figure 9 is a flowchart showing the processing performed by the speech recognition program (language decoder).

Figure 10 is a diagram showing the speech recognition program (language decoder) of Figure 7 wherein automatic disfluency deletion or automatic punctuating is not performed because only a general-purpose language model is used.

Figure 11 is a diagram showing the speech recognition program (language decoder) of Figure 7 wherein the general-purpose language model and the punctuation language model are used to perform automatic punctuating.

Figure 12 is a diagram showing the speech recognition program (language decoder) of Figure 7 wherein the general-purpose language model and the disfluency language model are used to perform automatic disfluency deletion.

Figure 13 is a diagram showing the speech recognition program (language decoder) of Figure 7 wherein the general-purpose language model, the disfluency language model, and the punctuation language model are used to perform automatic
20 disfluency deletion and automatic punctuating.

DETAILED DESCRIPTION OF THE INVENTION

The preferred embodiments of the present invention will now be described while referring to the accompanying drawings. Figure 1 is a diagram showing the configuration of a speech recognition apparatus 1 according to a first embodiment.

As shown in Figure 1, the speech recognition apparatus 1 can include a CPU 10, which further can include a micro processor, memory and peripheral circuits; an input device 12, including a mouse, a keyboard and a speech input device 120, such as a microphone; a display device 14, such as a CRT display; and a storage device 16, such as an HDD (Hard Disk Drive), a DVD (Digital Versatile Disc) device, or a CD (Compact Disk) device.

The speech recognition apparatus 1 employs a hardware configuration wherein the speech input device 120 is added to a common computer, and executes a recorded speech recognition program 2a that is supplied on a storage medium 18, such as a DVD, a CD-ROM or a CD-R. That is, the speech recognition apparatus 1 recognizes speech (speech data) that is received via the speech input device 120 and is transformed into a digital form, generates text data from which disfluencies are automatically removed, and stores the text data on the storage medium 18 that is loaded into the storage device 16, or displays the text data on the display device 14.

Figure 2 is a diagram showing the structure of the speech recognition program 2a. As shown in Figure 2, the speech recognition program 2a includes an acoustic processor 20, a language decoder 22, an acoustic model 30, a language model 32a, a baseform pool 34a, and an ON/OFF controller 36 for the language model 32a. The language model 32a also includes a general-purpose language model 320 and a disfluency language model (topic) 322.

The acoustic processor 20 performs a frequency analysis process and a feature extraction process, such as a Fourier transformation, for the speech data that is input at the speech input device 120, transforms the speech data into a feature vector, and outputs the feature vector to the language decoder 22. The acoustic model 30 outputs to the language decoder 22, acoustic model data wherein the appearance probability (a

numerical value is increased as the probability becomes greater that a phoneme sequence represents a word sequence) of the feature vector is entered for each phoneme by using HMMs (Hidden Markov Models). The baseform pool 34a includes a general-purpose portion that corresponds to the general-purpose language model 320 of the language model 32a and a disfluency pronunciation (e.g., "er") that corresponds to the disfluency language model 322. The baseform pool 34a writes the pronunciation of each disfluency word using a phonetic symbol that is correlated with the acoustic model 30.

While N (e.g., N = 3) consecutive words are employed for the appearance probability of a word or a word sequence that is generally required for speech recognition, the general-purpose language model 320 of the language model 32a is written in table form in correlation with the appearance probability. The size of the general-purpose language model 320 is normally about 60 MB. The appearance probability of the combination of words when N = 3 is defined as the probability one succeeding word of two consecutive words will appear, and in the general-purpose language model 320, a disfluency is not regarded as a word to be recognized.

The disfluency language model 322 includes a word set (in the example where N = 3, "this", "is" and "er") that includes a disfluency and a word set (e.g., "you" and "know") that is related to a colloquial expression unique to spontaneous speech. Both of these word sets are prepared by scanning text that is obtained by transcribing spontaneous speech in order to learn the disfluency language model 322. A disfluency is also one of the expressions unique to spontaneous speech. While the second word set is not directly related to disfluency detection, this effectively improves the precision with which spontaneous speech is recognized, because most of the general-purpose language models 320 are obtained by learning "written words" from, for example, newspapers.

Figs. 3A and 3B are diagrams showing the operation of the speech recognition program 2a (the language decoder 22) in Figure 2. In Figure 3A, speech recognition (topic OFF) is performed using only the general-purpose language model 320; the

disfluency language model 322 is not used. In Figure 3B, speech recognition (topic ON) is performed by using both the general-purpose language model 320 and the disfluency language model 322.

The language decoder 22 decodes the feature vector received from the acoustic processor 20, and outputs, to the display device 14, or the storage device 16, text data constituting a word sequence (W in equation (2)) for which the maximum probability is calculated by the acoustic model 30, the baseform pool 34a, and the language model 32a. The text data is then displayed on the display device 14, or stored in the storage device 16. As shown in Figures 3A and 3B, depending on whether the disfluency language model 322 is used, the language decoder 22 automatically detects and deletes a disfluency in a manner that will be described later while referring to Figures 4 to 6.

The ON/OFF controller 36 controls the ON/OFF state of the disfluency language model 322 and its use. Although various control methods are available, the easiest method to use is to determine the ON/OFF state of the disfluency language model 322 in accordance with a user's instruction. To control the ON/OFF state in an automated fashion, however, language decoding need only be performed twice for one part of the speech data for a user, i.e. when the disfluency language model 322 is ON and when it is OFF. The obtained scores (likelihoods) may be compared to determine the one that is the most appropriate, so that the ON/OFF control can be exercised. This determination can be performed each time the disfluency language model 322 is used; however, since CPU 10 processing time is required to perform language decoding twice, it is preferable that the determination be performed at the time of user registration or enrollment.

In equation (2), $P(S|W)$ is provided by the acoustic model 30 and the baseform pool 34a. $P(W)$ is provided by the language model 32a. By using equation (1), $P(W)$ is obtained by the weighted averaging of the general-purpose language model 320 and the disfluency language model 322 in accordance with the value for λ . When the value for λ is changed, not only the ON/OFF state of the disfluency language model 322 can

be controlled, but also the selection of the general-language model 320 or the disfluency language model 322. The importance level of the selected model also can be controlled. Generally, a system designer sets a value that, by evaluating results obtained by experiment, is assumed to be optimal; however, this value also can be adjusted by experimental decoding at the time enrollment is accomplished.

$$W' = \operatorname{argmax}_w P(W|S) = \operatorname{argmax}_w P(S|W)P(W) \dots (2)$$

where S denotes a feature vector (S_1, S_2, \dots, S_k), W denotes a word sequence (W_1, W_2, \dots, W_l) and W' denotes a word sequence (W'_1, W'_2, \dots, W'_l).

The operation of the language decoder 22 will now be described in detail while referring to Figures 4 to 6 and by using a 3-gram as an exemplary N-gram. Figure 4 is a flowchart showing the speech recognition processing performed by the speech recognition program 2a (language decoder 22) of Figure 2. Figure 5 is a diagram showing the operation of the speech recognition program 2a (language decoder 22) that does not perform automatic disfluency deleting (topic OFF) because only the general-purpose language model 320 of Figure 2 is used. Figure 6 is a diagram showing the operation of the speech recognition program 2a (language decoder 22) that performs the automatic deletion of disfluencies (topic ON) by using both the general-purpose language model 320 and the disfluency language model 322 of Figure 2.

As shown in Figure 4, at step 100 (S100 in Figure 4; this step notation method applies hereinafter), the acoustic processor 20 transforms input speech "kore ga ehon" into feature vectors which can be output to the language decoder 22. At step 102 (S102), as shown in Figures 5 and 6, the language decoder 22 outputs the received feature vectors (S_1, S_2, \dots, S_k) to the acoustic model 30. The acoustic model 30 then prepares acoustic data obtained by adding the appearance probabilities to phoneme sequences (/koegaehon/ and /koregaehon/) that may correspond to the feature vectors, and returns the acoustic data to the language decoder 22.

At step 104 (S104), the language decoder 22 determines whether the disfluency automatic deletion function has been set in accordance with a user's instruction or by experimental decoding at the time of enrollment, i.e., whether the disfluency language model 322 (topic) is ON. When the automatic deletion function has not been set (the topic is OFF), the value held by λ is set to 1 and program control advances to step 110 (S110). When the automatic deletion function has been set, the value held by λ is set to 0 or 1, as has been determined in advance, and program control is shifted to step 120 (S120).

When the automatic deletion of disfluencies has not been set, at step 110 (S110), as shown in Figure 5, the language decoder 22 refers only to the general-purpose portion (general-purpose baseform pool) of the baseform pool 34a to transform the phoneme sequences (/koegaehon/ and /koregaehon/) into word sequences (koe/ga/ehon and kore/ga/ehon), and determines that the word sequences are choices, while not regarding disfluencies as choices. That is, the portion /ehon/ is transformed into a common word or a set of common words that is pertinent to the baseform /ehon/. It should be noted that this transformation is performed using only the probability supplied by the acoustic model 30.

At step 112 (S112), as shown in Figure 5, the language decoder 22 refers to only the general-purpose language model 320 to renew the appearance probabilities assigned to the word sequence choices that are obtained at step 110 (S110). In the example in Figure 5, as a result of this processing, the probability assigned to word sequence "kore/ga/ehon" is 0.02, and the probability assigned to word sequence "koe/ga/ehon" is 0.01.

To automatically delete a disfluency, at step 120 (S120), as shown in Figure 6, the language decoder 22 refers to both the general-purpose portion of the baseform pool 34a and the disfluency portion (disfluency baseform pool) written in phonetic expression of the disfluency, and transforms the phoneme sequences (/koegaehon/ and /koregaehon/) into word sequences (e.g., koe/ga/ehon, koe/ga/<er>/hon and kore/ga/<er>/hon) while including disfluencies as word choices. It should be noted that

er is enclosed by angle brackets, <>, to indicate that it is a disfluency, an efficient measure for use when displaying text data from which disfluencies are to be automatically deleted.

At step 122 (S122), as shown in Figure 6, the language decoder 22 refers to the general-purpose language model 320 and the disfluency language model 322 by weighting them in accordance with equation (1), wherein $\lambda \neq 1$, and renews the appearance probabilities assigned to the word sequence choices obtained at S110. In the example in Figure 6, as a result of this processing, the probability assigned to word sequence "kore/ga/<er>/hon" is 0.02, and the probability assigned to each of the remaining word sequences is 0.01.

At step 124 (S124), as shown in Figures 5 and 6, the language decoder 22 outputs text data that represents the speech recognition result, the word sequence for which the highest probability is obtained by performing a calculation or renewing at S112 or S122. In the example in Figure 6, "kore/ga/<er>/hon" is selected, and although in Figure 6 <er> is displayed, the text data also can be displayed after <er> has actually been deleted. This applies for the following examples, and even when <er> appears in a display, this includes a case wherein it is not actually displayed.

The processing performed when a speaker inputs as speech "korega" "e" "hon" at the speech input device 120 (Figure 1) of the speech recognition apparatus 1 will now be separately described for a case wherein the disfluency language model 322 is used and for a case wherein it is not used. When the disfluency language model 322 is not used, the acoustic processor 20 processes speech data for the speech that is input, and outputs to the language decoder 22, the feature vectors that describe the speech sounds. The language decoder 22, as shown in Figures 4 and 5, processes the feature vectors received from the acoustic processor 20 by using only the acoustic model 30 and the general-purpose language model 320, and evaluates the probabilities obtained for the acoustic model 30 and the language model 32a to identity "korega" as "kore/ga" and "e" "hon" as "ehon". That is, the sound "e", for which the highest probability is provided by referring to the acoustic model 30 and the language model 32a, is selected

from the common word combinations, and as a result, in this case text data "kore/ga/ehon" is output.

When both the general-purpose language model 320 and the disfluency language model 322 are employed, the speech data is processed and the feature vectors that describe the sounds of the speech are output to the language decoder 22. Then, as shown in Figures 4 and 6, the language decoder 22 employs the acoustic model 30, the general-purpose language model 320, and the disfluency language model 322 to process the feature vectors received from the acoustic processor 20, evaluates the probabilities obtained by referring to the acoustic model 30 and the language model 32a, and identifies "korega" as "kore/ga" and "e" "hon" as "<er>" and "hon". That is, in accordance with the context, the word having the highest probability is determined for sound "e" by using the probability obtained from the language model 32a, while taking into account not only the common word "e", but also the disfluency "<er>". In the state in Figure 3, when the two word sequences "kore/ga/ehon/hatsumeii" and "kore/ga/<er>/hon/hatsumeii" are compared, it is more easily understood that the second word sequence is assigned a higher probability.

The disfluency language model 322 is constituted by a pronunciation dictionary to which a mark indicating a disfluency is allocated, and requires a much smaller size (e.g., 2 MB) than the above described task for automatic disfluency deletion.

Further, since for the automatic disfluency deleting method of this embodiment the disfluency language model 322 need merely be added as a topic, the automatic disfluency deletion function can be added without changing the general-purpose conventional system.

In addition, since the ON/OFF control does not require that the program be reset, the user can easily control the ON/OFF state of the automatic disfluency deletion function by selecting the disfluency language model 322 for this operation.

Furthermore, the disfluency language model 322 also can be used with a topic that is especially prepared for another field, such as "computer", "soccer" or "baseball". For example, a topic for inserting marks, such as punctuation, also can be provided. An

explanation will now be given, as a second embodiment, of an example wherein both the topic for disfluency deletion and the topic for punctuation insertion are provided.

Since the basic configuration of a speech recognition apparatus according to a second embodiment is the same as that of the speech recognition apparatus 1 for the first embodiment, no explanation for it will be given. Figure 7 is a diagram showing the arrangement of a speech recognition program 2b according to the second embodiment. Since the basic functions of the speech recognition program 2b are the same as those of the speech recognition program 2a in the first embodiment, the same reference numerals as are used for the speech recognition program 2a in the first embodiment are also used to denote corresponding components in Figure 7.

The speech recognition program 2b differs from the speech recognition program 2a in that a language model 32b includes a punctuation language model 323 in addition to a general-purpose language model 320 and a disfluency language model 322. The speech recognition program 2b also differs from the speech recognition program 2a in that a baseform pool 34b corresponds to the punctuation language model 323, and includes a punctuation portion for detecting a blank (pause; corresponding to a punctuation ("," or ",")) in the acoustic model data. The contents of the baseform pool 34b, which are not shown in Figure 7, will be described later.

When automatic punctuating is to be performed, the language decoder 22 regards a pause in the speech as a word. When automatic punctuating is not to be performed, the language decoder 22 does not regard the pause of the speech as a word.

Since the functions of the general-purpose language model 320 and the disfluency language model 322 of the language model 32b are the same as those for the first embodiment, only the punctuation language model 323 will be described. The punctuation language model 323 is a topic especially prepared for punctuation insertion. While a combination of three continuous words is employed for the appearance probability of a specific word/word sequence that is required for punctuation insertion, the punctuation language model 323 is represented in table form

in correlation with the appearance probability.

The punctuation language model 323 can be regarded as a general-purpose language model 320, which is specially prepared in order to automatically insert punctuation into portions that are determined as pauses by the baseform pool 34b, while regarding punctuation marks as words, and for which the volume of data is reduced.

The punctuation language model 323 includes words that are selected based on the amount of mutual information for the punctuation class, e.g., following words that are positioned immediately before the punctuation marks. The following enumerated words are the upper twenty words, and Hm indicates the value of the volume of the mutual information in the punctuation class. Since the punctuation language model 323 limits the information required for punctuation insertion, generally, the amount of data can be about 1/100 to 1/1000 of that of the general-purpose language model 320.

Words in the punctuation language model 323 can include:

Hm	words
275.111	iru
197.166	da
160.223	shita
159.425	desu
152.889	ha
137.400	shi
137.164	ne
129.855	de
112.604	aru
103.377	ga
79.751	masu
73.160	ka
66.952	shikashi
65.562	ori

63.930	node
63.078	mashita
62.469	imasu
59.100	daga
49.474	nai
48.714	deha

5

Figures 8A to 8D are diagrams showing the operation of the speech recognition program 2b (language decoder 22) of Figure 7. In Figure 8A, only the general-purpose language model 320 is used for speech recognition; in Figure 8B, the general-purpose language model 320 and the punctuation language model 323 are employed; in Figure 8C, the general-purpose language model 320 and the disfluency language model 322 are employed; and in Figure 8D, the general-purpose language model 320, the disfluency language model 322, and the punctuation language model 323 are employed.

The language decoder 22 decodes the feature vectors received from an acoustic processor 20, and outputs to a display device 14 or a storage device 16, text data constituting a word sequence (W' in equation (2)) for which the highest probability is obtained by the acoustic model 30, the baseform pool 34b, and the language model 32b. The text data is then displayed on the display device 14 or stored in the storage device 16.

As shown in Figures 8A to 8D, depending on whether the disfluency language model 322 and the punctuation language model 323 are used, the language decoder 22 automatically detects and deletes disfluencies and insert punctuations in a manner that will be described later while referring to Figures 9 to 13.

In equation (2), $P(S|W)$ is provided by the acoustic model 30 and the baseform pool 34b, and $P(W)$ is provided by the language model 32b. By using equation (1), $P(W)$ is obtained by weighted averaging of the general-purpose language model 320, the disfluency language model 322, and the punctuation language model 323 in

accordance with the value of λ . When the value of λ is changed, not only the ON/OFF states of the disfluency language model 322 and the punctuation language model 323 can be controlled, but also the selection of the general-language model 320, the disfluency language model 322, or the punctuation language model 323. Additionally, the importance level of the selected model can be controlled. Generally, a system designer sets a value which is assumed to be optimal from results obtained by experimentation; however, this value also can be adjusted by experimental decoding at the time of enrollment.

The operation of the language decoder 22 will now be described in detail while referring to Figures 9 to 13 and by using a 3-gram as an exemplary N-gram. Figure 9 is a flowchart showing the speech recognition processing performed by the speech recognition program 2b (language decoder 22) of Figure 7. Figure 10 is a diagram showing the operation of the speech recognition program 2b (language decoder 22) that does not perform the automatic deletion of disfluencies and automatic punctuating (two topics in the OFF state) because only the general-purpose language model 320 of Figure 7 is used. Figure 11 is a diagram showing the operation of the speech recognition program 2b (language decoder 22) that performs automatic punctuating (punctuation topic in the ON state) by using the general-purpose language model 320 and the punctuation language model 323 of Figure 7. Figure 12 is a diagram showing the operation of the speech recognition program 2b (language decoder 22) that performs automatic disfluency deletion (disfluency topic in the ON state) by using the general-purpose language model 320 and the disfluency language model 322 of Figure 7. Figure 13 is a diagram showing the operation of the speech recognition program 2b (language decoder 22) that performs automatic disfluency deletion (disfluency topic in the ON state) and automatic punctuating (punctuation topic in the ON state) by using the general-purpose language model 320, the disfluency language model 322, and the punctuation language model 323 of Figure 7.

As shown in Figure 9, at step 200 (S200 in Figure 9; this step notation method applies hereinafter), the acoustic processor 20 transforms the input speech "kore ga

ehon" into feature vectors, and outputs them to the language decoder 22. At step 202 (S202), as shown in Figures 10 to 13, the language decoder 22 outputs the received feature vectors (S_1, S_2, \dots, S_k) to the acoustic model 30. The acoustic model 30 prepares acoustic data obtained by adding the appearance probabilities to phoneme sequences (/koegaehon/ and /koregaehon/) that may correspond to the feature vectors, and returns the acoustic data to the language decoder 22.

At step 204 (S204), the language decoder 22 determines whether the automatic disfluency deletion function is set in accordance with an instruction issued by a user or by experimental decoding at the time of enrollment, i.e., whether the disfluency language model 322 (disfluency topic) is ON. When the automatic deletion function is not set (the disfluency topic is OFF), a value held by λ is set to 1, and program control advances to step 208 (S208). When the automatic deletion function is set, the value held by λ is set to 0 or 1, as determined in advance, and program control is shifted to step 206 (S206).

At 206 (S206), the language decoder 22 determines whether the automatic punctuating function is set in accordance with an instruction issued by the user or by experimental decoding at the time of enrollment, i.e., whether the punctuation language model 323 (punctuation topic) is ON. When the automatic punctuating function is not set (the punctuation topic is OFF), the value held by λ is set to 1, and program control advances to step 220 (S220). When the automatic punctuating function is set, the value held by λ is set to 0 or 1, as determined in advance, and program control is shifted to step 210 (S210). The process at step 210 is performed when both the disfluency topic and the punctuation topic are ON, and the process at step 220 is performed when the disfluency topic is ON.

At step 208 (S208), as well as at step 206 (S206), the language decoder 22 determines whether the punctuation language model 323 (punctuation topic) is ON. When the automatic punctuating function is not set (the punctuation topic is OFF), the value held by λ is set to 1, and program control advances to step 240 (S240). When the automatic punctuating function is set, the value held by λ is set to 0 or 1, as

determined in advance, and program control is shifted to step 230 (S230). The process at step 230 is performed when the punctuation topic is ON, and the process at step 240 is performed when both the disfluency topic and the punctuation topic are OFF.

At step 240 (S240), as shown in Figure 10, the language decoder 22 refers to only the general-purpose portion (general-purpose baseform pool) of the baseform pool 34b to transform the phoneme sequences (e.g., /koegaehon/ and /koregaehon/) into word sequences (e.g., koe/ga/,/ehon and kore/ga/,/ehon), and determines the word sequences are choices, without regarding disfluencies and pauses as choices. That is, the portion /ehon/ is transformed into a common word or a set of common words that is pertinent to the baseform /ehon/. It should be noted that this transformation is performed by using only the probability provided by the acoustic model 30.

At step 242 (S242), as shown in Figure 10, the language decoder 22 refers only to the general-purpose language model 320 to renew the appearance probabilities of the word sequence choices that are obtained at step 220 (S220). In the example in Figure 10, as a result of this processing, the probability assigned to word sequence "kore/ga/,/ehon" is 0.02, and the probability assigned to word sequence "koe/ga/,/ehon" is 0.01.

At step 230 (S230), as shown in Figure 11, the language decoder 22 refers to both the general-purpose baseform pool and the punctuation portion (punctuation topic) for pause detection, which are included in the baseform pool 34b, to transform the phoneme sequences (e.g., /koegaehon/ and /koregaehon/) into word sequences (e.g., koe/ga/ehon, /kore/ga/e/hon/, /koe/ga/,/ehon/, /kore/ga/,/e/hon), while regarding pauses as words.

At step 232 (S232), as shown in Figure 11, the language decoder 22 refers to the general-purpose language model 320 and the punctuation language model 323 by weighted averaging of them in accordance with equation (1), wherein $\lambda \neq 1$, and renews the appearance probabilities assigned to the word sequence choices that are obtained at step 230 (S230). In the example in Figure 11, as a result of this processing, the probability assigned to word sequence "/kore/ga/,/e/hon/" is 0.02, and the probability

assigned to each of the other word sequences is 0.01.

At step 220 (S220), as shown in Figure 12, the language decoder 22 refers to both the general-purpose baseform pool of the baseform pool 34b and the disfluency portion (disfluency topic) thereof that is written expressed in phonetic symbols, and transforms the phoneme sequences (e.g., /koegaehon/ and /koregaehon/) into word sequences (e.g., koe/ga/,/ehon, kore/ga/,/ehon and kore/ga/,/er>/hon), while regarding disfluencies as word choices. As when the topic is OFF, this transformation is performed by using only the probability provided by the acoustic model 30.

At step 222 (S222), as shown in Figure 12, the language decoder 22 refers to the general-purpose language model 320 and the disfluency language model 322 by weighting them in accordance with equation (1), wherein $\lambda \neq 1$, and renews the appearance probabilities assigned to the word sequence choices obtained at S220. In the example in Figure 12, as a result of this processing, the probability assigned to word sequence "kore/ga/,/er>/hon" is 0.02, and the probability assigned to each of the remaining word sequences is 0.01.

At step 210 (S210), as shown in Figure 13, the language decoder 22 refers to the general-purpose baseform pool, the disfluency portion (disfluency topic) and the punctuation portion (punctuation topic) included in the baseform pool 34b, and transforms the phoneme sequences (e.g., /koegaehon/ and /koregaehon/) into word sequences (e.g., koe/ga/ehon, kore/ga/er>/hon and kore/ga/,/er>/hon), while regarding disfluencies and pauses as word choices.

At step 212 (S212), as shown in Figure 13, the language decoder 22 refers to the general-purpose language model 320, the disfluency language model 322 and the punctuation language model 323 by weighting them in accordance with equation (1), wherein $\lambda \neq 1$, and renews the appearance probabilities assigned to the word sequence choices obtained at S210. In the example in Figure 13, as a result of this processing, the probability assigned to word sequence "kore/ga/,/er>/hon" is 0.02, and the probability assigned to each of the remaining word sequences is 0.01.

At step 224 (S224), as shown in Figures 10 to 13, the language decoder 22

sequentially outputs as text data representing the speech recognition results, the word sequences for which the assigned appearance probabilities are the highest as the renewing results obtained at S212, S222, S232 and S242.

The operation of the speech recognition apparatus 1 (Figure 1 or 7) for the second embodiment will now be explained while referring to Figures 8A to 8D. When the disfluency language model 322 and the punctuation language model 323 are not employed ($\lambda = 1$), and when a speaker inputs, at the speech input device 120 (Figure 1) of the speech recognition apparatus 1, speech consisting of, for example, "korega" "pause (indicates no sound; this applies hereinafter)" "ten" "e" "honhatsumeino" "pause" "youten", as shown in Figure 8A, the acoustic processor 20 outputs the feature vectors that represent this speech.

As shown in Figures 7 and 10, the language decoder 22 (Figure 2) employs only the acoustic model 30 and the general-purpose language model 320 to process the feature vectors received from the acoustic processor 20, and evaluates the probabilities supplied by the acoustic model 30 and the general-purpose language model 320. Thus, the language decoder 22 identifies "korega" as "kore/ga", "ten" following "pause" as comma ",", "e" "hon" as "e-hon", and "hatsumeino" as "hatsu-mei no". Further, since "ten" and "maru" do not follow the "pause" following "honhatsumeino", the language decoder 22 identifies the succeeding "youten" as "you-ten", and outputs, as the identification results, the text data "kore ga, e-hon-hatsu-mei no you-ten".

As shown in Figure 8B, when the disfluency language model 322 is not employed ($\lambda = 1$), and when at the speech input device 120 (Figure 1) of the speech recognition apparatus 1 a speaker inputs "korega" "pause" "e" "honhatsumeino" "pause" "youten", the acoustic processor 20 outputs, to the language decoder 22, the feature vectors that represent the phonemes of the speech.

As shown in Figures 7 and 11, the language decoder 22 (Figure 2) employs the acoustic model 30, the general-purpose language model 320, and the punctuation language model 323 to process the feature vectors received from the acoustic processor 20, and evaluates the probability obtained from the acoustic model 30 and

the probability obtained from the general-purpose language model 320 and the punctuation language model 323. Thus, the language decoder 22 identifies "korega" as "kore/ga", "pause" following "ga" as comma ",", "e" "hon" as "e-hon", and "hatsumeino" as "hatsu-mei no". Further, since "pause" follows "hatsumeino" and normally no punctuation follows "no", the language decoder 22 identifies the succeeding "youten" as "you-ten", not inserting punctuation at the "pause" portion, and outputs, as the identification results, the text data "kore ga, e-hon-hatsu-mei no you-ten".

As shown in Figure 8C, when the punctuation language model 323 is not employed ($\lambda = 1$), and when at the speech input device 120 (Figure 1) of the speech recognition apparatus 1 a speaker inputs "korega" "pause" "ten" "e" "honhatsumeino" "pause" "youten", the acoustic processor 20 outputs, to the language decoder 22, the feature vectors that represent the phonemes of the speech.

As shown in Figures 7 and 12, the language decoder 22 (Figure 2) employs the general-purpose language model 320 and the disfluency language model 322 to process the feature vectors received from the acoustic processor 20, and evaluates the probability obtained from the acoustic model 30, and the probability obtained from the general-purpose language model 320 and the disfluency language model 322. Thus, the language decoder 22 identifies "korega" as "kore/ga", "ten" following "pause" as comma ",", and "e" "hon" as "<er>" "hon". That is, a word that in context has a higher probability is determined by referring to the probabilities obtained from the language model 32b, while taking into account that not only the normal word "e" but also the disfluency "<er>" may be applied for sound "e". Further, since neither "ten" nor "maru" follows the "pause" following "hatsumeino", the language decoder 22 identifies the succeeding "youten" as "you-ten", and outputs, as the identification results, the text data "kore ga, <er> hon-hatsu-mei no you-ten". In the condition shown in Figure 8C, when the two word sequences "kore/ga/,ehon/hatsumeino" and "kore/ga/,<er>/hon/hatsumeino" are compared, it is easily understood that of the two, the second word sequence is the most probable.

As shown in Figure 8D, when the disfluency language model 322 and the

punctuation language model 323 are employed ($\lambda \neq 1$), and when, unlike the above case, at the speech input device 120 (Figure 1) of the speech recognition apparatus 1 a speaker inputs "korega" "pause" "e" "honhatsumeino" "pause" "youten" excluding "ten", the acoustic processor 20 outputs, to the language decoder 22, the feature vectors that represent the phonemes of the speech.

As shown in Figures 7 and 12, the language decoder 22 (Figure 2) employs the acoustic model 30, the general-purpose language model 320, the disfluency language model 322, and the punctuation language model 323 to process the feature vectors received from the acoustic processor 20, and evaluates the probability obtained from the acoustic model 30, and the probability provided by the general-purpose language model 320, the disfluency language model 322, and the punctuation language model 323. Thus, the language decoder 22 identifies "korega" as "kore/ga", "pause", which follows "ga" in "korega", as comma ",", and "e" "hon" as "<er>" "hon". That is, a word that in context has a higher probability is determined by referring to the probabilities obtained from the language model 32b, while taking into account not only that the normal word "e" but also the disfluency "<er>" may be applied for the sound "e". In the condition shown in Figure 8D, when the two word sequences "kore/ga/,ehon/hatsumeino" and "kore/ga/,<er>/hon/hatsumeino" are compared, it is easily understood that of the two, the second word sequence is most probable. In addition, since "pause" follows "hatsumeino" and generally no punctuation follows "no", the language decoder 22 does not insert punctuation into the "pause" portion. Similar to the case wherein the punctuation language model 323 is not employed, the language decoder 22 identifies the input speech accurately as "korega, <er> hon-hatsu-mei no you-ten", and outputs it as the text data that is the identification result.

According to the second embodiment, the following effects are obtained in addition to those obtained by the first embodiment. The automatic punctuating function can be added to the conventional general-purpose system without changing the system substantially. Further, without resetting the program, the user can turn on or off the automatic punctuating function merely by selecting or not selecting the automatic

punctuating topic.

Furthermore, when a change in the frequency of the punctuation inserted depends on the content of a document, according to the punctuation insertion method of the embodiment, the appearance frequency of the punctuation can be easily
5 changed by adjusting the weight used in the linear interpolation with the general-purpose language model 320.

The present invention has been explained by referring to two embodiments. These embodiments are provided for the Japanese language. However, the disfluency processing can be applied not only to Japanese, but also to other foreign languages, such as English. Thus, an example will be given wherein a disfluency affects English speech recognition results. "She devised remedial measures." is used as an example sentence, which is translated into "kano-jo ha zen-go-saku wo kanga-e da-shita" in Japanese. Assuming that a speaker voices the disfluency <uh> between "She" and "devised", <uh> "devised" may be erroneously recognized as "advised". When the
10 embodiments of the present invention are employed for such an English sentence, and when the disfluency language model 322 is OFF, the system recognizes "She advised remedial measures." When the disfluency language model 322 is ON, the system recognizes "She devised remedial measures."

As described above, according to the present invention, when a general-purpose
20 language model in addition to a vocabulary/language model especially prepared for disfluencies and punctuation marks is employed, disfluencies can be deleted from appropriate positions in sentences; and, marks, such as punctuation marks, can be inserted.